Introduction

In the post human genome sequencing era, multiple genome-wide profiling experimental paradigms, including array-based comparative genomic hybridization and expression profiling and high throughput proteomics, have been developed and used extensively by the scientific cancer community. Integration of the existing knowledgebase of individual genes, their mRNA and protein products, and their biochemical and genetic interactions (heretofore collectively referred to as gene or genomic annotation) with the analysis of such experiments is fundamental to interpreting this data correctly so that viable, new hypotheses may be generated. This genomic annotation, which resides in hundreds of individual databases, represents an enormous fund of knowledge that can be leveraged with experimentally- and clinically-derived data by physician-scientists to gain meaningful insights into pathophysiological processes governing tumorigenesis. However, individual researchers commonly report that finding this information is incredibly daunting and time consuming since the data is scattered across the web. Underscoring this fact, the 2004 inaugural issue of *Nucleic Acids Research* catalogued over 500 individual molecular biology databases, representing an increase of over 150 databases from last year (Galperin, 2004).

Frequently, biomedical scientists rely on bioinformaticians to facilitate this task, and they are most often successful in extracting meaningful information from a handful of databases. However, they, too, find that the task of data integration across many of these databases is very difficult because of several reasons including but not limited to:

- There are no standards which effectively encapsulate genomic annotation data as a whole. Thus, annotation data is available in various ad hoc text and binary formats from different data sources.
- There is no single, common identifier, or reusable common data element (CDE), for a gene or its mRNA or protein product. This results in individual silos of information, which cannot be interlinked syntactically.
- Controlled vocabularies are infrequently used, making the task of semantic integration impossible.

To address this and other data integration problems, the National Cancer Institute began the **ca**ncer Biomedical **I**nformatics **G**rid (caBIG) initiative "to expedite the cancer research communities' access to key bioinformatics platforms [through] an integrating biomedical informatics infrastructure that integrates diverse data types and supports interoperable analytic tools" (Buetow, 2004). In brief, the caBIG project, which includes three domain and two cross-cutting workspaces, aims to deploy an interconnected grid of biomedical *in silico* services. These services can be broadly categorized into three major categories:

- Data services- These include data sources which will serve out primary experimental data sets (e.g. microarray or proteomics data) as well as tissue-related information (e.g. pathology or histological data) and clinical data (e.g. laboratory values or demographic information).
- Analysis services- These include software applications aimed at analyzing experimental, tissue-based, and clinical data either separately or concurrently.
- Annotation services- These include data sources which will serve out genomic annotation data sets.

Although many other use cases are possible, one broadly envisioned use case is that data services will be queried to acquire desired experimental, pathology, and clinical data, which will be piped to analysis services. Results from analysis services will be the input to annotation services,

resulting in a final integrated data set for the physician-scientist to examine. This use case illustrates several requirements from the design perspective. First, one or more CDEs representing gene identifiers are required to execute such a use case across all three services. Second, CDEs must be standardized across annotation services so that all the available annotation for a particular gene may be syntactically and semantically integrated.
Therefore, this example and other such use cases clearly define the scope of the data integration problem facing the biomedical cancer community and the caBIG consortium. Because this problem is too complex as a whole, it must be tackled in manageable pieces. Specifically, the Genomic Annotation Special Interest Group within the Integrated Cancer Research (ICR) Workspace has been charged with integrating all the available collections of biomedical genomic information, which will be represented as annotation services on the grid. It is the objective of this document to make recommendations on how to facilitate this task.

Recommendations

The National Cancer Institute Center for Bioinformatics (NCICB) has developed a number of core infrastructure tools which may be used to integrate data across the different ICR projects. First, the **ca**ncer **D**ata **S**tandards **R**epository was created to store Unified Modeling Language (UML) models as interrelated objects comprised of ISO/IEC 11179 compliant CDEs. In this system, an individual CDE is defined by a data element concept and a value domain. A second tool, the Enterprise Vocabulary Service (EVS), contains controlled terminology required to define the data element concept and the value domain. The caDSR and the EVS may be utilized as follows to facilitate data integration across the different ICR projects:

- Each project must describe its objects using a UML model.
- This model's objects must contain ISO/IEC 11179 consistent CDEs.
- This model must be imported into the caDSR.
- The same CDEs representing gene identifiers should be used across projects.

Although extreme exceptions exist, in such a system, it is feasible to relate data models by joining them through shared CDEs. Because the same CDEs are utilized across two projects, the join is syntactically sound (i.e. a LocusLinkId CDE used in one model is equal to a LocusLinkId CDE used in another model), and because these CDEs are defined using the same terminology, the join is also semantically appropriate in the overwhelming majority of cases.
However, a fundamental problem with using the same CDE is that there is no overarching identifier that represents a gene or its mRNA or protein product across all publicly available databases. This is exemplified by the fact that NCBI and Ensembl, two of the largest bioinformatics database producers use different identifiers between each other (e.g. Entrez Gene ID versus Ensembl Gene Identifier) and even internally (e.g. LocusLink ID versus UniGene ID). Therefore, we propose the following recommendations pertaining to the creation and usage of CDEs representing a gene or its mRNA or protein product to the ICR community:

- A list of required CDEs representing a gene or its mRNA or protein product should be gathered by examining the data models of the current ICR projects and by scrutinizing potential future use cases.
- This list of CDEs should either be reused from the existing CDEs in the caDSR or should be created by the ICR community
- Each ICR project's data model should utilize **at least** one or more of these defined CDEs.
- Because the goal is not to be restrictive, if data models in the future cannot reasonably accommodate one of the existing CDEs, additional CDEs may be added to this dynamic

list. The Architecture Workspace, Vocabulary/CDE Workspace, and genome annotation subject matter experts (potentially the Genome Annotation SIG members) should oversee the addition of such CDEs.

To initiate the process of defining this list of CDEs, we suggest encapsulating the following gene/mRNA/protein identifiers as CDEs:

- Accession Number (GenBank, UniProt, SwissProt, TrEMBL, DDBJ)
- NCBI GI
- UniGene ID
- LocusLink ID
- RefSeq DNA ID (NT_...)
- RefSeq mRNA ID (NM_..., XM_...)
- RefSeq Protein ID (NP_..., XP_...)
- EMBL ID
- Ensembl Gene ID (ENSG…)
- Ensembl Transcript ID (ENST…)
- Ensembl Protein ID (ENSP…)
- SwissProt ID
- TrEMBL ID
- HUGO ID
- UniProt ID
- PIR ID
- PIR NREF ID

The Genome Annotation SIG should review this document and examine the list of potential CDEs for additions and/or deletions. Furthermore, through active teleconference, online, or newsgroup discussions, this SIG should arrive at a final CDE list in the upcoming weeks so that ICR projects may proceed to develop Silver level compliant applications which require no modification of the data model to achieve Gold level compatibility in the future.

References

Buetow, K. (2004). cancer Biomedical Informatics Grid. http://cabig.nci.nih.gov.

Galperin, M. Y. (2004). The Molecular Biology Database Collection: 2004 update. Nucleic Acids Res *32 Database issue,* D3-22.